Practical uses for POWER in HTC

Tristan Suerink, IT Architect
NIKHEF

OpenPOWER Summit Europe
RAI Centre | Amsterdam
October 3-4, 2018

Join the Conversation #OpenPOWERSummit
• Interaction and structure of all elementary particles and fields

– Accelerator physics - CERN
– Astroparticle physics

• Knowledge and technology transfer
• (Astro-) Particle physics collaboration

  – Science program with University partners
  – Infrastructure Nikhef institute
    • Technical & engineering support
      – Mechanical workshop
      – Electronic workshop
    • Large computing infrastructure (T1)
    • Long term strategy & commitment

<table>
<thead>
<tr>
<th>Staff Type</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permanent Staff</td>
<td>71</td>
</tr>
<tr>
<td>PhD+PostDoc</td>
<td>120</td>
</tr>
<tr>
<td>Technical/engineer</td>
<td>70</td>
</tr>
<tr>
<td>Support</td>
<td>28</td>
</tr>
</tbody>
</table>
LHC programme - Dutch involvement

Long term LHC program and roadmap investment funding (phase1 & phase2)

Excellent scientific programme

Contribution to CERN: 53 MCHF - 4.7%
European Strategy for Astroparticle Physics
APPEC–Astroparticle Physics European Consortium

APP portfolio @ Nikhef

- Pierre Auger - cosmic rays
- Xenon1T - Dark Matter
- Adv VIRGO - Gravitational Waves
- KM3NeT - neutrino detection
Veel instrumenten ter beschikking voor onderzoek – deeltjesfysica, astrodeeltjesfysica, kosmologie

Nikhef technical groups
Tier1 centre - Steadily growing GRID infrastructure -

- over 9000 cores
- over 6 Petabyte of storage
- state-of-the-art 6 terabit/s network

Available to non-HEP clients (~8%)
HPC vs HTC

• High Performance Computing:
  – “classic” Supercomputers
  – Running a few big jobs over multiple nodes
  – ±2GB ram per core
  – Infiniband/Omnipath interconnect
  – Shared filesystem

• High Throughput Computing:
  – Grid computing
  – Running multiple jobs per machine from multiple users
  – ±8GB ram per core
  – Highspeed Ethernet
  – No shared filesystem
Typical questions/reactions from others

• AMD has a killer CPU, so why looking further?!?  
• GPU/FPGA are the future, so you won’t need CPU’s!  
• ARM will be marketleader within 5 years!  
• Intel will fix everything within no-time  
• I don’t need more performance  
• POWER is only useful for banks and big enterprise  
• POWER costs too much!  
• I don’t trust anything else than Intel!  
• More cores is always better!
Why have we looked into POWER?
Why have we looked into POWER?

• Launch of OpenPOWER in 2015
• More GHz
• More PCI-e lanes
• I/O optimized
• Configurable cache coherency
• Looked promising
• Something different compared with x86
What happened next?

• So we bought our first POWER machine back in 2015!
• IBM S822L with 2x 8 cores and 128GB ram
HP DL380 gen 9 vs IBM S822L
LXC Busybox containers with a dhcp client
Started all containers in parallel
HP crashed after several minutes trying…
IBM did it in 3 minutes

1000 LXC containers
Let's generate some packets!
Let's generate some packets!

witchport mode: access
learning mode: Enabled
forwarding mode: inherited cut-through
flowmetry sampling: Disabled  TCs: N\A
flowmetry threshold: Disabled  TCs: N\A
flowmetry threshold level: N\A
last clearing of "show interface" counters : Never
30 seconds ingress rate: 336232 packets 281008 bits/sec, 42029028876 bytes
30 seconds egress rate: 536 bits/sec, 67 bytes/sec, 1 packets/sec

<table>
<thead>
<tr>
<th></th>
<th>packets</th>
<th>unicast packets</th>
<th>multicast packets</th>
<th>broadcast packets</th>
<th>bytes</th>
<th>error packets</th>
<th>discard packets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rx</td>
<td>1605952388</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1605949922</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>153</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2313</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14440771848932</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>34156043</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tx</th>
<th>packets</th>
<th>unicast packets</th>
<th>multicast packets</th>
<th>broadcast packets</th>
<th>bytes</th>
<th>error packets</th>
<th>discard packets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1262</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
High speed network test

- 1x IBM S922L - POWER9 - 2x 8 core
- 5x Mellanox Connect X5 PCI-e gen4 dual-port
- 8x 100Gbit/s usable
- Bonded as one LACP trunk
- Hitting kernel limits at ±350Gbit/s
- Only using 25% CPU usage

- Goal was to reach the 800Gbit/s…
Data Transfer Nodes (DTN)

- Storage machines tuned for long distance data transfers
- Capable of handling multiple transfer protocols
- Helping scientists to get the data where they want
- Used in a science network
- Firewall destroyers ;-) 
- Completely different storage design needed
Our DTN requirements

- 8MiB/s per TiB read
- 4MiB/s per TiB write
- 2 reads and 1 write per 100TiB at the same time
- RAID6-like redundancy with maximum of 15 drives
- GridFTP/xrootd/webdav
- Running CentOS
- Redundant connections to external enclosures
- Manageable for a real Sysadmin
Typical setup
Typical setup specs

- 1x Fujitsu RX2530M4 with 2x Intel 4110 and 40Gbit/s NIC
- 1x Seagate AssuredSAN 4004
- 2x hardware raid controllers with SAS3 connectors
- 56x 8TB NLSAS drives
- ±350TiB netto storage
- Uses in total 5U cabinet space
- ±36Gbit/s throughput measured at NIC
- Maximum cache mirroring 6GB/s
- Using 8TB drives, 3.6GB/s is needed

- So we can grow using 12TB drives!
Let’s try something fun!
Let’s see if we can reach the 100Gbit/s

1x IBM S822L - POWER8 - 2x 8 core
2x Seagate AssuredSAN 4004
2x LSI 9405W-16e
2x 350TiB netto storage
8x SAS3 connections
2x 100Gbit/s ethernet
No SSD’s!

Let’s compare with:
Fujitsu RX2530M4 - 2x Intel Xeon 4110 Silver
Dell R7415 - 1x AMD EPYC 7451
The results!

Reads and Writes combined

<table>
<thead>
<tr>
<th></th>
<th>Intel Xeon</th>
<th>AMD EPYC</th>
<th>POWER8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R1</strong></td>
<td>2.764</td>
<td>1.777</td>
<td>3.260</td>
</tr>
<tr>
<td><strong>W1</strong></td>
<td>3.660</td>
<td>2.226</td>
<td>6.284</td>
</tr>
<tr>
<td><strong>R1+W1</strong></td>
<td>3.991</td>
<td>3.991</td>
<td>5.072</td>
</tr>
<tr>
<td><strong>R2+W1</strong></td>
<td>4.502</td>
<td>4.402</td>
<td></td>
</tr>
<tr>
<td><strong>R8+W4</strong></td>
<td>3.909</td>
<td>4.764</td>
<td>10.135</td>
</tr>
<tr>
<td><strong>R16+W8</strong></td>
<td>3.224</td>
<td>3.587</td>
<td>8.945</td>
</tr>
</tbody>
</table>

**8.945**
What to choose?

OR?
• A lot of libs aren’t optimized for ppc64le yet
• There is not an easy way to test code on ppc64le?
• Small differences between IBM POWER and OpenPOWER
• At the moment I can only buy POWER via IBM in NL
• OpenCAPI and OpenPOWER? Where is it?
Questions for the audience

• Why doesn’t the CentOS/RHEL installer work out of the box?
• Why isn’t there an iso for FreeBSD and Gentoo?
• Why doesn’t petitboot/OPAL include tools like storcli?
• Why not send the screen to all hvc/tty interfaces by default?
• Where are the OpenCAPI NICs?
• OpenCAPI vs Gen-Z? What if?
• 16GB Hugepages doesn’t work at the moment. Why?
• POWER8 and 9 differences in docs are unclear.
• What’s the difference between ppc64le and ppc64el?
Questions?
Test environment for DTN test

- OS: Ubuntu 18.04
- Test suite: http://web.grid.sara.nl/acceptance_test/
- Used the dd test from this suite
- 350TiB of testfiles
- Re-used the same HBAs and SAS cables
- All firmwares on the latest stable level
- Hyperthreading turned off
- Optimal memory configuration for every platform
- More info: https://wiki.nikhef.nl/grid/100g_Storage_BOX